

Does Financial Speech Need Continual Learning Regularization? A Study of EWC and mini-KFAC on Earnings Call ASR

Thai Quoc On

Department of Computer Science
Columbia University, NYC, NY, USA
tqo2101@columbia.edu

Homayoon Beigi

Mechanical Engineering Dept. and
Electrical Engineering Dept.
Columbia University, NYC, NY, USA
and Recognition Technologies, Inc.
South Salem, NY, USA
homayoon.beigi@columbia.edu

Abstract

Earnings call transcription presents a realistic continual learning scenario where Automatic Speech Recognition (ASR) systems are updated with new recordings but risk catastrophic forgetting of previously learned financial speech patterns. Adapting Whisper-small from a general-domain speech corpus to the Earnings22 benchmark reveals the opposite: all evaluated methods exhibit improvement on held-out financial speech from the same domain during adaptation rather than forgetting. This task complementarity makes regularization counterproductive. Nevertheless, when regularization is warranted, a layer-wise mini-Kronecker-Factored Approximate Curvature (mini-KFAC) approximation seems to achieve lower word error rates (WERs) and more stable optimization than standard diagonal Elastic Weight Consolidation (EWC). The results suggest that task complementarity should be assessed before applying continual learning constraints, and prefer mini-KFAC EWC over diagonal EWC when needed.

1 Introduction

Financial speech recognition systems face a continual learning challenge in which models must be updated as new data arrives, raising concerns about catastrophic forgetting during sequential fine-tuning (Kirkpatrick et al., 2017; Rusu et al., 2016). However, when tasks share acoustic and linguistic structure, adaptation to one partition may improve performance on another, a phenomenon referred to here as task complementarity. Whether catastrophic forgetting actually occurs in financial speech Automatic Speech Recognition (ASR) remains underexplored. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) constrains parameter updates via Fisher information, but diagonal approximations ignore parameter correlations that may be important in large transformer-based ASR models such as Whisper (Radford et al.,

2023). Although McInerney et al. (2021) showed that Kronecker-factored EWC substantially reduces forgetting relative to diagonal EWC, this question has not been examined in financial speech.

This work adapts Whisper-small from a TEDLIUM-derived corpus (Rousseau et al., 2012; Gandhi, 2023) to the Earnings22 benchmark (Rio et al., 2022), comparing unregularized fine-tuning, diagonal EWC, and a memory-efficient Kronecker-factored variant: mini-Kronecker-Factored Approximate Curvature EWC (mini-KFAC EWC). The central finding is that all methods show statistically significant negative forgetting, with performance on held-out financial speech improving during adaptation rather than degrading, within this single-domain setting. These findings contribute two practical guidelines to the continual learning literature: task similarity assessment should precede any regularization decision, and Kronecker-factored approximations are preferable to diagonal Fisher methods when constraints are necessary.

2 Related Work

Continual learning studies how to learn sequential tasks without forgetting prior knowledge. Specifically, EWC (Kirkpatrick et al., 2017) addresses this by penalizing parameter drift using Fisher information, while replay and architectural expansion methods offer alternative approaches (Rebuffi et al., 2017; Rusu et al., 2016; Zenke et al., 2017; Serra et al., 2018).

ASR continual learning remains less studied than vision (Vander Eeck and Van Hamme, 2022; Plantinga et al., 2023), and financial speech benchmarks such as Earnings22 (Rio et al., 2022) are underexplored in this context.

Kronecker-Factored Approximate Curvature (KFAC) approximates the Fisher information layer-wise, as $F \approx G \otimes A$ where A is input covariance, G is gradient covariance, and $\otimes$ denotes the

Kronecker product (Martens and Grosse, 2015), capturing parameter dependencies more effectively than diagonal methods while remaining scalable (Martens and Grosse, 2015). Prior work has shown the effectiveness of Fisher-based regularization in other domains. For example, McInerney et al. (2021) reported that Kronecker-factored EWC reduced forgetting by 51%, compared with 27% for diagonal EWC, on BERT-based medical entity linking. Additionally, Ahadzi et al. (2025) applied EWC to Whisper-small for children’s ASR and achieved a 5.21% relative Word Error Rate (WER) reduction over unregularized fine-tuning. However, neither work examines Kronecker-factored approximations for ASR nor the financial speech domain, and neither reports the task complementarity regime.

3 Experimental Setup

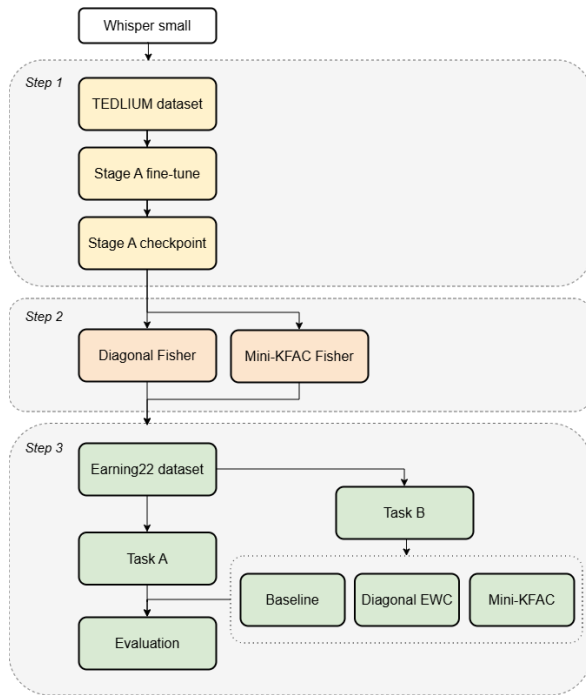


Figure 1: The architecture of the continual learning experimental pipeline.

Model and Data: Whisper-small (Radford et al., 2023) (244M parameters), a transformer encoder-decoder model pre-trained on large-scale weakly supervised speech data, is used. The experimental pipeline in Figure 1 proceeds in three steps and mimics a realistic incremental deployment, in which a model is first adapted to a general-domain corpus and then fine-tuned on financial speech, with retention measured on a disjoint

partition of the target domain. Particularly, in **Step 1 (Stage A)**, the model is adapted on the training split of TEDLIUM-derived speech corpus (Rousseau et al., 2012; Gandhi, 2023), a large corpus of aligned TED talk speech, to provide a general-domain initialization. In **Step 2 (Fisher Estimation)**, the Fisher statistics are estimated from Stage A batches using both diagonal and Kronecker-factored (mini-KFAC) approximations, capturing parameter importance before financial-domain adaptation begins. In **Step 3 (Stage B)**, the model is fine-tuned on the Earnings22 data (Rio et al., 2022), which is partitioned into two disjoint subsets through MD5 hashing of file identifiers: **Task B**, the target partition used for Stage B fine-tuning, and **Task A**, a held-out retention partition on which the model is never trained, used to evaluate whether prior performance is preserved during the Stage B adaptation. Since each file identifier corresponds to a distinct earnings call recording, this partitioning guarantees non-overlapping speaker groups across Task A and Task B. All three methods: unregularized baseline, diagonal EWC, and mini-KFAC EWC are evaluated on both Task A and Task B partitions. Task construction rationale and hyperparameter details are in Appendix A and B, and the full code implementation is available at <https://anonymous.4open.science/r/continual-asr-finance-508B/>.

Methods:

- **Baseline:** Unregularized fine-tuning.
- **Diagonal EWC:** Parameter-wise ℓ_2 regularization (Kirkpatrick et al., 2017): $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ASR}} + \frac{\lambda}{2} \sum_i F_i (\theta_i - \theta_i^*)^2$, where $\mathcal{L}_{\text{ASR}}$ is the ASR loss, θ_i and θ_i^* are the current and Stage A parameters, F_i is the diagonal Fisher estimate for parameter i , and λ is the regularization strength.
- **Mini-KFAC EWC:** Layer-wise Kronecker-factored Fisher regularization with per-layer adaptive scaling $s_\ell = \sqrt{\overline{\text{Tr}}(A_\ell) \cdot \overline{\text{Tr}}(G_\ell)}$, where A_ℓ and G_ℓ are the Kronecker factors (input activation and output gradient covariances) for layer ℓ , and $\overline{\text{Tr}}(\cdot)$ is the mean trace (trace divided by matrix dimension); s_ℓ is clipped to $[1.0, 5.0]$.

Mini-KFAC was selected due to the computational cost of full KFAC. This approach was termed mini-KFAC to reflect three memory-efficient approximations relative to the original formulation

Table 1: Results across five seeds (mean $\pm$ std). Task B WER is the best validation WER on the fine-tuning partition. Task A WER is evaluated on the best Task B checkpoint. Forgetting is the change in Task A WER relative to pre-B; negative values indicate improvement. Superscripts indicate statistical significance. $\dagger$ denotes a one-sample, one-tailed t -test showing that forgetting is significantly below zero ($p < 0.05$), see Appendix D.

Method	Task B WER (%)	Task A WER (%)	Forgetting (%)
Baseline	31.94 $\pm$ 1.73	20.68 $\pm$ 1.63	-4.87 $\pm$ 2.47 †
EWC (Diagonal)	35.23 $\pm$ 4.91	21.88 $\pm$ 2.08	-3.66 $\pm$ 1.95 †
Mini-KFAC EWC	34.26 $\pm$ 3.57	21.52 $\pm$ 2.17	-4.02 $\pm$ 2.41 †

(Martens and Grosse, 2015): restriction to linear layers with dimension ≤ 2048 , row subsampling of activation and gradient matrices to at most 100 rows per batch, and CPU-side accumulation of the Kronecker factors. Additional rationale details are provided in Appendix B. In this set-up, a unified layer-wise regularization scheme is applied across all continual learning methods such that encoder layers receive stronger regularization ($\lambda_{\text{enc}} = 10,000$) than decoder layers ($\lambda_{\text{dec}} = 2500$) given the encoder’s greater acoustic specificity (Vander Eeck and Van Hamme, 2022). The lower global scale for mini-KFAC against EWC ($\lambda_{\text{scale}} = 48$ vs. 10,000) compensates for the larger magnitude of matrix quadratic-form penalties relative to scalar per-parameter terms (Appendix B). Moreover, coefficients were selected empirically to keep the EWC regularization influential but non-dominant, with regularization strength ramped linearly from zero over the first 50 steps (Appendix E).

Evaluation Metrics: Three metrics are reported: (1) The Task B WER, which is the best word error rate on the Task B partition during fine-tuning; (2) The Task A WER, signifying the word error rate on the Task A retention partition after fine-tuning completes; and (3) The forgetting rate, defined as the change in the Task A WER before and after the Stage B fine-tuning, i.e. Forgetting = $\text{WER}_A^{\text{final}} - \text{WER}_A^{\text{pre-B}}$, where pre-B is the Task A WER evaluated immediately after the Stage A training and before any Stage B updates. Negative forgetting indicates that Task A performance improved during Stage B training rather than degraded. All results are reported as mean $\pm$ standard deviation across five seeds (3, 17, 24, 35, and 50).

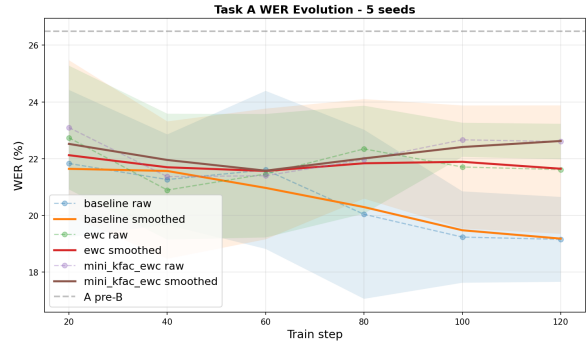


Figure 2: Task A WER evolution during Stage B fine-tuning.

4 Results

4.1 Finding 1: Task Complementarity Makes Regularization Counterproductive

All three methods exhibit statistically significant negative forgetting: Task A WER improves during Stage B fine-tuning for every method (one-sample t -tests: baseline $t = -4.41$, $p = 0.006$; EWC $t = -4.20$, $p = 0.007$; mini-KFAC $t = -3.73$, $p = 0.010$; all one-tailed, $df = 4$). The stronger the regularization constraint, the less improvement is permitted: baseline forgetting is -4.87% , mini-KFAC -4.02% , and diagonal EWC -3.66% . This pattern indicates that earnings call speech shares acoustic and linguistic patterns across both partitions, so regularization prevents beneficial co-adaptation rather than blocking genuine interference, making it counterproductive in this setting.

4.2 Finding 2: Unregularized Fine-tuning Remains Superior

Despite the advantage of mini-KFAC over diagonal EWC, baseline fine-tuning achieves lower WER than both regularized methods on Task B (31.94% vs. 35.23% for EWC and 34.26% for mini-KFAC). Although pairwise t -tests are not statistically significant at $n = 5$, the effect sizes for Task B WER are large (baseline vs. EWC: $d = 0.89$; baseline vs. mini-KFAC: $d = 0.83$), indicating a practically meaningful pattern that more seeds would likely confirm. This baseline advantage is visible in Figure 3, where the baseline achieves the lowest WER throughout training.

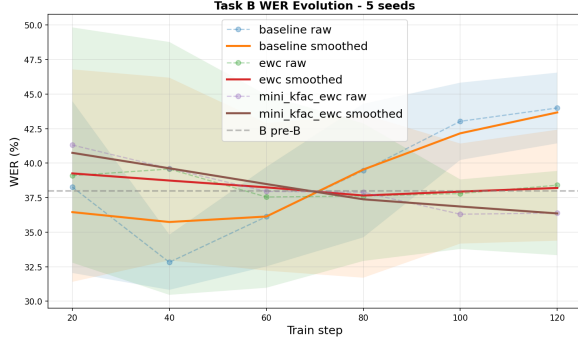


Figure 3: Task B WER evolution during Stage B fine-tuning.

4.3 Finding 3: Mini-KFAC Shows Consistent but Statistically Modest Advantages over Diagonal EWC

When regularization is applied, mini-KFAC EWC achieves lower WER than diagonal EWC across all metrics (Table 1): a 0.97% Task B WER reduction (34.26% vs. 35.23%) and a 0.36% Task A WER reduction (21.52% vs. 21.88%). Mini-KFAC also shows lower variance across seeds (Task B std: 3.57 vs. 4.91), indicating more stable optimization. Figures 2 and 3 show that mini-KFAC converges faster and to a lower WER than diagonal EWC on both tasks across all seeds. However, Welch’s two-sample t -tests indicate that these differences are not statistically significant at $n = 5$ seeds (Task A: $t = -0.27$, $p = 0.80$; Task B: $t = -0.36$, $p = 0.73$; effect sizes $d \approx 0.17$ – 0.23 , small; see Appendix D). The consistent directional advantage of mini-KFAC across all metrics and seeds suggests a genuine but modest benefit of Kronecker factorization: by capturing within-layer parameter correlations, mini-KFAC allows more selective constraints than purely diagonal penalties, though larger-scale experiments are needed to confirm this statistically.

5 Discussion

Interpreting the absence of forgetting: The lack of forgetting is not merely a consequence of the task construction. Although Task A and Task B come from the same dataset, meaningful interference between disjoint partitions would still be expected to cause degradation. Instead, the consistent improvement on Task A suggests that positive transfer dominates in this setting. This highlights the importance of assessing task relevance before applying continual learning regularization, as strong transfer can render such constraints unnecessary or

even detrimental.

Scope of the “forgetting” framing: The phrase “negative forgetting” is used in this work as a label for the signed WER change defined in Section 3, not as a claim of resistance to forgetting under genuine domain shift. Task A and Task B are both Earnings22 partitions from the same financial-speech distribution, so the Task A improvement reflects fine-tuning generalizing across speakers within one domain, rather than the classical retention-adaptation trade-off studied under larger domain shifts. This within-domain design was chosen deliberately (Appendix A): it isolates regularization effects from confounding domain shift, and represents the realistic setting of incremental updates from the same data distribution. Because the two tasks are closely related, the setup approximates a best case for avoiding forgetting, and the results show that regularization is unnecessary and counterproductive here. The results say nothing about what happens under larger domain shift, a question discussed in Appendix F.

Sources of task complementarity: The two corpora draw from non-overlapping speaker populations: TEDLIUM encompasses 228 named TED conference speakers (Gandhi, 2023), distinct from global corporate executives and financial analysts in Earnings22 (Rio et al., 2022). Both consist of prepared or semi-prepared professional English speech with controlled recording conditions that provide cleaner signal than far-field or telephone-quality speech (Gandhi, 2023; Rio et al., 2022). These shared properties are consistent with the encoder representations acquired during Stage A generalizing constructively to Earnings22, and offer a plausible account of the negative forgetting observed across all methods.

Mini-KFAC outperformance over diagonal EWC: Diagonal Fisher treats parameters independently, producing overly restrictive constraints in high dimensions. Combined with large but effective regularization strength, this makes diagonal EWC rigid. In contrast, Mini-KFAC uses a Kronecker factorization that captures layer-wise structure, allowing more flexible updates that better preserve second-order geometry. Its layer-wise scaling further corrects Fisher magnitude imbalance across the network, which diagonal EWC cannot address.

Practical recommendations: Practitioners should include evaluating forgetting before applying continual learning regularization as stable or negative forgetting indicates task complementar-

ity and reduces the need for additional constraints. When catastrophic forgetting is observed, mini-KFAC EWC is preferable to diagonal EWC. Moreover, careful tuning of regularization strength is also important, as overly restrictive constraints can limit adaptation and degrade performance, while overly loose or small values can provide insufficient penalty to preserve prior knowledge, making the regularization ineffective.

6 Conclusion

In this study, a comparative evaluation of Fisher-based continual learning methods is presented for financial speech recognition, using the Whisper-small model. The presented results suggest that preserving layer-wise parameter interactions through Kronecker-factored approximations may provide more effective regularization than diagonal Fisher approximations in this setting. Furthermore, all methods exhibit negative forgetting, with Task A performance improving throughout the Stage B fine-tuning process. This observation suggests strong task complementarity in the financial speech domain and, in this within-domain setting, contrasts with the commonly expected retention–adaptation trade-off in continual learning, where adaptation to a new task often degrades performance on previously learned tasks. These findings establish a practical guideline: task similarity should be assessed before applying regularization, and when regularization is warranted, mini-KFAC EWC might be the better choice.

7 Limitations

This study uses a controlled two-partition Earnings22 setup and reports results over five random seeds. While five seeds yield statistical significance of negative forgetting ($p < 0.05$ across all methods, with two of three methods at $p < 0.01$), pairwise comparisons between regularization strategies remain underpowered at this scale (effect sizes $d \leq 0.23$ for mini-KFAC vs. diagonal EWC). These tests also rest on a normality assumption that is difficult to verify at $n = 5$ and Appendix D discusses this limitation and notes bootstrap-based alternatives as a more robust next step. Larger seed counts and longer task sequences remain important directions for future work to confirm the relative advantage of Kronecker-factored regularization. Because Task A and Task B are both partitions of the same Earnings22 corpus (Appendix A), the nega-

tive forgetting observed here is a within-domain result and should not be read as evidence that regularization is unnecessary under larger domain shifts. Additionally, scenarios where regularization may be beneficial are discussed in Appendix F.

References

- Edem Ahadzi, Vishwanath Pratap Singh, Tomi Kinnunen, and Ville Hautamaki. 2025. Continuous learning for children’s asr: Overcoming catastrophic forgetting with elastic weight consolidation and synaptic intelligence. *arXiv preprint arXiv:2505.20216*.
- M. Bisani and H. Ney. 2004. [Bootstrap estimates for confidence intervals in asr performance evaluation](#). In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–409.
- Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jetté. 2021. [Earnings-21: A Practical Benchmark for ASR in the Wild](#). In *Inter-speech 2021*, pages 3465–3469.
- Sanchit Gandhi. 2023. TEDLIUM data. <https://huggingface.co/datasets/sanchit-gandhi/tedlium-data>. Hugging Face dataset.
- L. Gillick and S.J. Cox. 1989. [Some statistical issues in the comparison of speech recognition algorithms](#). In *International Conference on Acoustics, Speech, and Signal Processing*,., pages 532–535 vol.1.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. [Overcoming catastrophic forgetting in neural networks](#). *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Zhe Liu and Fuchun Peng. 2020. [Statistical testing on asr performance via blockwise bootstrap](#). *Preprint*, arXiv:1912.09508.
- James Martens and Roger Grosse. 2015. [Optimizing neural networks with kronecker-factored approximate curvature](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417, Lille, France. PMLR.
- Denis Jered McInerney, Luyang Kong, Kristjan Arumae, Byron Wallace, and Parminder Bhatia. 2021. [Kronecker factorization for preventing catastrophic forgetting in large-scale medical entity linking](#). *Preprint*, arXiv:2111.06012.
- Peter Plantinga, Jaekwon Yoo, and Chandra Dhir. 2023. Continual learning for end-to-end asr by averaging domain experts. *arXiv preprint arXiv:2305.09681*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. 2022. [Earnings-22: A practical benchmark for accents in the wild](#). *Preprint*, arXiv:2203.15591.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. [TED-LIUM: an automatic speech recognition dedicated corpus](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 125–129, Istanbul, Turkey. European Language Resources Association (ELRA).

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR.

Steven Vander Eeckt and Hugo Van Hamme. 2022. Continual learning for monolingual end-to-end automatic speech recognition. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 459–463. IEEE.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. Pmlr.

A Task Construction Rationale

Task A and Task B are disjoint Earnings22 partitions with different speakers, and recordings, enabling controlled evaluation of retention within the same financial-speech domain. This study uses within-domain partitions rather than cross-domain tasks for two reasons. First, this isolates the effect of Fisher-based regularization from the confounding influence of large distributional shifts and any forgetting measured here is directly attributable to parameter drift rather than domain mismatch. Second, this represents a realistic deployment scenario:

production earnings-call systems are updated incrementally as new quarterly recordings become available, where successive batches are drawn from the same underlying financial-speech distribution. This controlled design provides a lower-bound study of catastrophic forgetting. Stronger domain shifts (e.g., Earnings22 followed by a medical or conversational corpus) may produce different regularization dynamics, and should be considered as an important direction for future work. Additionally, Earnings22 is used rather than Earnings21 (Del Rio et al., 2021) to avoid manual segmentation steps that could confound the continual learning analysis.

B Hyperparameter Details

All experiments were conducted on a single NVIDIA L4 GPU. Hyperparameters and settings were selected to fit within the available GPU memory and runtime budget while maintaining stable optimization. Full training configuration:

- Learning rate (Stage B): 3×10^{-6}
- Training steps: 120 steps (Stage A: 100, Stage B: 20), where one step corresponds to a single forward and backward pass on one mini-batch
- Evaluation interval: every 20 training steps (10 optimizer updates)
- Gradient accumulation: 2 steps per optimizer update (effective batch size 8)
- Batch size: 4 samples per training step
- Early stopping patience: 6 evaluations
- Warmup ratio: 10% of total steps
- EWC ramp-up: 50 steps (linearly scale λ from 0 to target)
- Fisher computation: 800 batches

Regularization coefficients:

- Diagonal EWC: $\lambda_{\text{base}} = 5000$, $\lambda_{\text{enc}} = 10,000$, $\lambda_{\text{dec}} = 2500$; global scale $\lambda_{\text{scale}} = 10,000$
- Mini-KFAC EWC: same base coefficients; global scale $\lambda_{\text{scale}} = 48$

Dataset partitioning: Earnings22 recordings are deterministically split into disjoint Task A and Task B subsets via MD5 hashing of file identifiers.

Stage B fine-tuning uses only Task B; Task A is never seen during training and is used solely to measure retention performance.

Regularization scale rationale: The large disparity between the diagonal EWC global scale ($\lambda_{\text{scale}} = 10,000$) and the mini-KFAC global scale ($\lambda_{\text{scale}} = 48$) is intentional and not a tuning artifact, where λ_{scale} is the global multiplier applied to the EWC penalty term before it is added to the ASR loss. The diagonal EWC penalty sums scalar Fisher estimates over individual parameters; each term is small. Specifically, the diagonal EWC penalty takes the form $\frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_i^*)^2$, where $F_i \in \mathbb{R}$ is the scalar diagonal Fisher estimate for parameter i , $\theta_i \in \mathbb{R}$ is the current value of parameter i , and $\theta_i^* \in \mathbb{R}$ is its value at the end of Stage A; each term is a scalar and the sum grows only linearly with the number of parameters. The mini-KFAC penalty instead computes a matrix quadratic form $\Delta W_\ell^\top (G_\ell \otimes A_\ell) \Delta W_\ell$ over the full weight matrix of each layer, which is naturally orders of magnitude larger in raw value, where $\Delta W_\ell = W_\ell - W_\ell^* \in \mathbb{R}^{p \times q}$ is the weight change matrix for layer ℓ (current minus Stage A weights), $G_\ell \in \mathbb{R}^{p \times p}$ is the output gradient covariance, $A_\ell \in \mathbb{R}^{q \times q}$ is the input activation covariance, $\otimes$ denotes the Kronecker product, and the resulting quadratic form aggregates correlations across all $p \times q$ parameters of the layer simultaneously rather than summing independent scalar terms. Both λ_{scale} coefficients were set so that the gradient ratio $\|\nabla_\theta \mathcal{L}_{\text{EWC}}\| / \|\nabla_\theta \mathcal{L}_{\text{ASR}}\|$ remains in a balanced, non-dominating range throughout Stage B training, where $\nabla_\theta \mathcal{L}_{\text{EWC}}$ is the gradient of the regularization term with respect to all model parameters θ , $\nabla_\theta \mathcal{L}_{\text{ASR}}$ is the corresponding gradient of the ASR loss, and a ratio well below 1 ensures the constraint guides but does not overwhelm task learning; this was verified empirically (Appendix E).

Mini-KFAC selection rationale: Computing the full Fisher matrix for a 244M-parameter model is computationally intractable. Mini-KFAC addresses this by approximating it layer-wise as a Kronecker product of input activation and output gradient covariances. This is computed using subsampled activations and gradient statistics extracted during GPU-based forward and backward passes, while Fisher accumulation is performed on the CPU to reduce GPU memory overhead. This hybrid design enables efficient estimation over approximately 800 batches within a single-GPU training budget, while still capturing inter-parameter

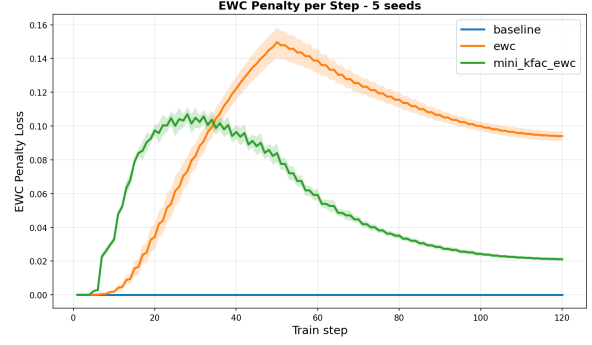


Figure 4: EWC penalty evolution during Stage B adaptation.

correlations that purely diagonal methods ignore.

C Training Dynamics Analysis

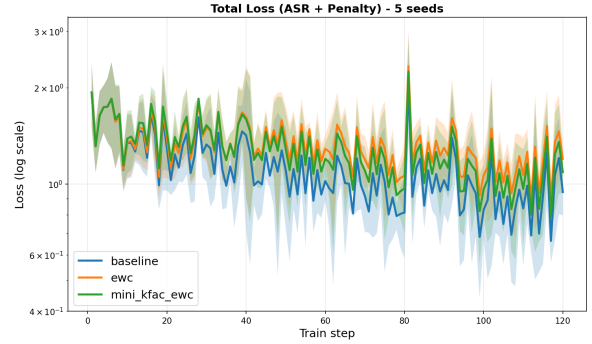


Figure 5: Total training loss ($\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ASR}} + \mathcal{L}_{\text{EWC}}$) during Stage B adaptation.



Figure 6: ASR loss component ($\mathcal{L}_{\text{ASR}}$) during Stage B adaptation.

Figure 4 shows the penalty dynamics during Stage B fine-tuning. Both EWC methods rise initially due to two compounding effects: a 50-step linear ramp-up of the global scaling coefficient λ , from 0 to its target value, and growing parameter drift ($\theta - \theta^*$) as the model adapts away from its Stage A weights. These cause the effective

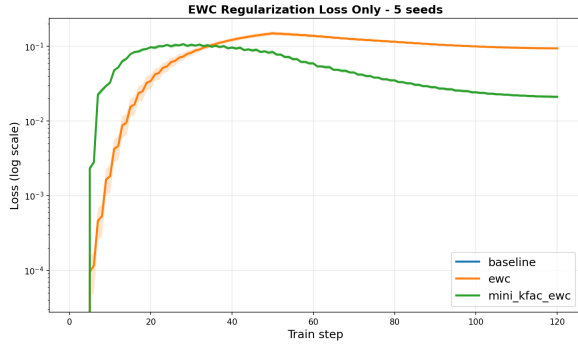


Figure 7: EWC regularization loss component ($\mathcal{L}_{EWC}$) during Stage B adaptation. The baseline shows zero EWC loss throughout, as no regularization is applied.

penalty to increase even before the ramp completes. Mini-KFAC rises faster and peaks earlier as its Kronecker-factored structure detects structured parameter drift more precisely, while its penalty then decreases as the model finds weight-space directions satisfying both Task B and the KFAC geometry. Diagonal EWC lacks this structure, accumulating penalty more uniformly and sustaining a higher, flatter trajectory throughout training.

The loss component breakdown (Figures 5, 6, and 7): The same ramp-up dynamic explains the initial rise in total loss, $\mathcal{L}_{\text{total}}$, for regularized methods: as λ scales up and parameter drift grows, the EWC term adds an increasing contribution to $\mathcal{L}_{\text{total}}$, constraining ASR improvement relative to the baseline. Mini-KFAC stabilizes at a lower, better-calibrated constraint level, allowing greater ASR progress while preserving retention, which explains its WER gains over diagonal EWC.

D Statistical Significance

Unit of analysis and test validity: The t -tests reported here rely on two assumptions: (1) the five WER values per method are approximately independent outcomes of repeated randomized training, and (2) they are approximately normally distributed. Both assumptions apply equally to the one-sample tests for negative forgetting (Table 3) and the pairwise comparisons between methods (Table 2), since both are computed from the same five per-seed WER values for each method. Both assumptions also differ from the concern raised in Gillick and Cox (1989), whose independence requirement addresses correlated word- or utterance-level errors within a single system’s output on a single fixed test set, rather than independence across five separately-trained systems.

(1) *Independence.* The five observations are approximately independent because each model is trained independently using a different random seed. Each seed determines the model initialization and dropout randomness, the batch order for both Stage A and Stage B training, which Stage A batches are used for Fisher estimation, and which rows are subsampled when estimating the mini-KFAC Kronecker factors (Appendix B). The evaluation partition is fixed and unshuffled, with Task A/B membership assigned deterministically from a hash of each file identifier rather than from the seed, so all five runs of a method are scored on the same utterances. Holding the evaluation set fixed removes evaluation-set variability and ensures that differences in WER primarily reflect differences induced by the randomized training procedure rather than differences in the data being evaluated. This supports treating the five WER measurements as approximately independent for both the one-sample and pairwise tests. It does not imply generalization beyond this particular validation partition, which is a separate question of external validity.

(2) *Normality.* The reported t -tests assume that the five per-seed WER values are approximately normally distributed. Although this assumption cannot be verified empirically at $n = 5$, its practical consequences differ between them. The one-sample tests for negative forgetting involve effect sizes that are large relative to the observed sampling variability (Table 3, $p < 0.05$ for all methods), so modest departures from normality are less likely to alter the qualitative conclusion than for the smaller pairwise differences. The pairwise comparisons between regularization strategies involve smaller observed differences, so the same unverifiable normality assumption is more consequential, and the tests correspondingly have limited power to detect true effects, consistent with the non-significant results in Table 2 despite directionally consistent trends across seeds. The findings therefore treat the pairwise comparisons as suggestive rather than conclusive, while placing more confidence in the one-sample negative-forgetting results. Bootstrap-based alternatives would help substantiate the pairwise comparisons further, since neither Bisani and Ney (2004)’s confidence intervals nor Liu and Peng (2020)’s blockwise bootstrap require a normality assumption, and both remain valid at small n .

Pairwise method comparisons: Table 2 reports Welch’s two-sample t -tests comparing all method pairs across three metrics (Task A WER, Task B WER, Forgetting), with Cohen’s d as a complementary effect size measure. No pairwise comparison reaches statistical significance at $n = 5$. The mini-KFAC vs. diagonal EWC differences are small in both absolute and effect-size terms ($d \leq 0.23$), confirming that stronger conclusions about relative regularizer quality require more seeds or longer task sequences. The baseline advantage over regularized methods, while not significant, exhibits large effect sizes on Task B WER ($d \approx 0.83$ – 0.89), suggesting a real pattern that larger experiments could confirm.

Table 2: Welch’s two-sample t -tests ($n = 5$ seeds per method, two-tailed). d = Cohen’s d (pooled std). No comparison is significant at $\alpha = 0.05$.

Comparison	Metric	Δ (%)	t	df	p	d
mini-KFAC vs EWC	Task A WER	−0.36	−0.27	8.0	0.796	−0.17
	Task B WER	−0.97	−0.36	7.3	0.731	−0.23
	Forgetting	−0.36	−0.26	7.7	0.802	−0.16
Baseline vs EWC	Task A WER	−1.20	−1.02	7.6	0.341	−0.64
	Task B WER	−3.29	−1.41	5.0	0.217	−0.89
	Forgetting	−1.21	−0.86	7.6	0.416	−0.54
Baseline vs mini-KFAC	Task A WER	−0.84	−0.69	7.4	0.510	−0.44
	Task B WER	−2.32	−1.31	5.8	0.241	−0.83
	Forgetting	−0.85	−0.55	8.0	0.597	−0.35

Negative forgetting (task complementarity):

To test whether forgetting is significantly negative which means whether Task A performance genuinely improves during Stage B, one-sample t -tests are applied against the null hypothesis of zero forgetting (H_0 : Forgetting ≥ 0), using one-tailed tests:

$$t = \frac{\bar{x}_{\text{forgetting}}}{s/\sqrt{n}}, \quad df = n - 1 = 4$$

Table 3: One-sample, one-tailed t -tests against H_0 : Forgetting ≥ 0 , $n = 5$ seeds per method, $df = 4$.

Method	t	p (one-tailed)	Sig.
Baseline	−4.41	0.006	$p < 0.01$
EWC (Diagonal)	−4.20	0.007	$p < 0.01$
Mini-KFAC EWC	−3.73	0.010	$p < 0.05$

All three methods show statistically significant negative forgetting ($p < 0.05$, one-tailed), confirming that task complementarity in financial speech is a robust empirical phenomenon and not an artifact of small sample variance.

E Gradient Flow Analysis

To monitor the interaction between the ASR objective and EWC regularization throughout Stage B training, gradient magnitude ratios are tracked:

$$\text{grad_ratio} = \frac{\|\nabla_{\theta} \mathcal{L}_{EWC}\|}{\|\nabla_{\theta} \mathcal{L}_{ASR}\| + \epsilon}$$

Regularized methods maintain stable non-zero ratios after the 50-step ramp-up, confirming that the EWC term contributes consistently to $\mathcal{L}_{\text{total}} = \mathcal{L}_{ASR} + \mathcal{L}_{EWC}$ rather than vanishing or dominating. The baseline has zero EWC gradient throughout, allowing unconstrained optimization toward the lowest ASR loss. Moreover, smooth ratio curves indicate stable optimization dynamics for all methods.

F When These Findings May Not Apply

These results are obtained in a setting where Task A and Task B are closely related financial-speech tasks and exhibit negative forgetting. Different conclusions may arise when tasks are less aligned. In particular, continual-learning regularization may be beneficial under three conditions: (1) substantial acoustic or linguistic mismatch between tasks, (2) strong catastrophic forgetting during sequential adaptation, or (3) limited Task B training data, where regularization can help preserve prior knowledge and reduce overfitting. These scenarios remain important directions for future evaluation.

G Future Directions

1. Task similarity assessment: Develop quantitative measures of task relatedness in financial speech (representation similarity, Fisher spectrum alignment, loss overlap) to predict when regularization is beneficial.
2. Adaptive regularization: Design dynamic λ weighting that decreases when Task A performance improves during Stage B adaptation, enabling more flexible transfer in complementary domains.
3. Selective layer-wise constraints: Investigate regularization of only layers with high task specificity through gradient attribution or attention sensitivity, rather than global Fisher penalties.

4. Longer financial sequences: Extend evaluation to multi-stage streams (for example, earnings calls across fiscal quarters, investor presentations, and regulatory transcripts) to determine whether complementarity persists.
5. Broader baselines: Compare Fisher-based methods against replay, adapter-based fine-tuning, and progressive architectures to situate mini-KFAC EWC within the broader continual learning landscape for financial ASR.